Detection of Hate Speech in Arabic

Subjects: Computer Science, Artificial Intelligence

Contributor: Malik Almaliki, Abdulqader M. Almars, Ibrahim Gad, El-Sayed Atlam

Hate speech towards a group or an individual based on their perceived identity, such as ethnicity, religion, or nationality, is widely and rapidly spreading on social media platforms. This causes harmful impacts on users of these platforms and the quality of online shared content. Fortunately, researchers have developed different machine learning algorithms to automatically detect hate speech on social media platforms. However, most of these algorithms focus on the detection of hate speech that appears in English. There is a lack of studies on the detection of hate speech in Arabic due to the language's complex nature.

Keywords: BERT ; hate speech ; Arabic ; detection ; classifier ; sentiment analysis

1. Introduction

Social media platforms, such as WhatsApp, Facebook, and Twitter, are widely used for exchanging and creating content. They provide users with a convenient and easy way to share information quickly and efficiently, making them a valuable source of information [1][2][3]. However, social media platforms can also be a means for disseminating offensive and harmful content. The propagation of unpleasant and harmful content on social media can have a significant damaging influence on the experience of users as well as the overall quality of online shared content [4]. Hate speech is an example of such harmful content which can be defined as speech that attacks or incites hatred against someone or something based on their perceived identity, such as ethnicity, religion, nationality, or sexual orientation [5][6][7][8]. According to a recent study conducted by the Anti-Defamation League (ADL), 41% of Americans had experienced online hate and harassment [9].

Therefore, building technologies that can automatically detect hate speech has become extremely critical. Fortunately, researchers in the fields of computer science and machine learning have developed algorithms that can automatically identify hate speech on social media platforms. These algorithms can help to mitigate the spread of this type of harmful content on these platforms. However, most of these algorithms focused on the detection of hate speech that appears in English [10][11][12][13][14]; there is a lack of studies on the detection of Arabic hate speech due to the language's complex nature. Prior studies on Arabic social media content have mostly concentrated on either recognizing vulgar or obscene language [15] or on the detection of hate speech that can be distinguished from it. The Arabic language is the main language in 6 of the top 11 countries with the highest social hostilities index, which evaluates crimes motivated in part by religion or race. This highlights the importance of addressing hate speech in Arabic, as this type of content can have serious negative consequences in communities [16].

The variety and complexity of Arabic morphology present certain difficulties for Arabic NLP researchers to detect hate speech on social media [17]. Dialectal Arabic is more frequently used in casual situations (e.g., social media platforms) than Modern Standard Arabic. Various dialects of Arabic exist within and between countries as well as among regions within the same country. There are no established grammar or spelling rules for dialectal Arabic, in contrast with Modern Standard Arabic [17]. It is common for similar-looking words to have different meanings in different dialects, which makes the language more ambiguous in general. For instance, the Arabic term "عافية" Afia in the Maghrebi Arabic language means "fire", whereas in Gulf Arabic it implies "health". The fact that Arabic has far fewer resources than English makes it more difficult. An Arabic hate vocabulary is one of the tools that is lacking, and it can be highly helpful in studies on cyber hate detection.

Furthermore, there are a number of difficulties in identifying hate and abusive speech on social media. Finding common patterns and trends in data is difficult due to the vast amount of diverse content that is uploaded to social media networks. Additionally, user-generated social network data includes noisy content that presents technological difficulties for text mining and linguistic analysis, such as incorrect grammar, misspelled words, internet slang, abbreviations, word lengthening, and text written in multi-lingual scripts. Finally, social network policies usually prohibit users from publishing

any unethical or unlawful content. Due to this, users post information that seems legitimate but very subtly escalates to the extremes of hate speech. As a result, it is difficult to develop tools that can detect hate speech automatically.

2. Detection Models of Hate Speech in Arabic

There have been several machine learning models proposed for identifying hate speech on social media platforms and other online communities. The topic of hate speech in English-language social media content has been studied in great detail. In [11], the researchers proposed a supervised method for identifying hate speech on Twitter. According to their findings, the use of supervised classifiers was found to be more effective in binary classification tasks than ternary classifiers. Burnap and Williams [18] have created a further binary classifier that uses a labeled dataset to distinguish between hateful and non-hateful tweets.

The textual characteristics of a message can be helpful in detecting hate speech. Using textual information from a user's tweets prior to them declaring support or opposition to ISIS, Magdy et al.'s classifier predicts whether a user supports ISIS or is against it [19]. Spatial and temporal features have also been used to identify hate speech. Jihadist et al. developed a model for identifying hate speech contents based on linguistic and temporal factors [20]. A method was developed by Mubarak et al. [15][21] for automatically building and growing a list of vocabulary words, which would subsequently be used to identify offensive tweets.

There has been some interest in using deep learning models to detect hate speech on social media platforms. Character n-grams are more accurate predictive variables for identifying racist and sexist tweets than word n-grams, according to Waseem and Hovy's theory. The researchers found that adding location information decreased performance, while adding gender as an additional variable only slightly improved it. Another research applied an LSTM-based classifier based on gradient-boosted decision trees (GBDTs) to detect hate speech. Compared to N-gram-based classifiers, this model outperformed them significantly [11].

Advanced models such as bidirectional encoder representations from transformers (BERT) have attracted the attention of scholars and practitioners [22][23][24]. BERT-large and BERT-base are two BERT models that were first presented by Devlin et al. [23] for automatically detecting hate speech in English. The proposed models were pre-trained based on quite substantial internet-extracted corpora. This results in enormous memory footprints and high computing demands. The proposed models make an effort to remedy some of the old models' flaws by enhancing either performance [25] or inference speed [22].

BERT models were also used to pre-train the Arabic language. As an example, Devlin et al. created a multilingual model that covers more than 100 languages, including Arabic [23]. According to Antoun et al., a BERT-based model named Arabert is pre-trained for Arabic content [14]. Around 24 terabytes of text were used for the model's pre-training. Similar to this, Abdul-Mageed et al. [19] trained an Arabic BERT model they called MARBERT using one billion tweets. Even though these models have been used to classify Arabic text, it is unclear if one is more effective than the other at detecting hate speech, or if the training process has affected their effectiveness.

A stacking BERT-based model for Arabic sentiment analysis was presented by Hasna et al. [26]. Transformer-based models were recently regarded as the most advanced model for several languages because of their excellent performance in sentiment analysis. However, Arabic sentiment analysis still needs to be more accurate. In this research, various BERT models are used to offer a stacking architecture for Arabic sentiment analysis. By combining various small, freely accessible datasets, a sizable Arabic sentiment analysis dataset is also produced. Experimental results show that the suggested approach is more accurate in classification than a single-model architecture. Muhammad et al. [27] suggested BERT semi-supervised learning of Arabic dialects.

The popularity of BERT led to more models supporting additional languages, including Arabic. BERT Models for Arabic Text Classification: A Systematic Review were proposed by [28]. Researchers and practitioners are paying more and more attention to bidirectional encoder representations from transformers (BERT), which has emerged as a crucial method for processing natural language. This method is successful for a variety of reasons, including its ability to predict words from context. It also has the ability to be pre-trained using a great deal of plain text data available online.

References

1. Chen, X.; Sin, S. ‘Misinformation? What of it?’ Motivations and individual differences in misinformation sharing on social media. Proc. Am. Soc. Inf. Sci. Technol. 2013, 50, 1–4.
2. Müller, K.; Schwarz, C. Fanning the Flames of Hate: Social Media and Hate Crime. J. Eur. Econ. Assoc. 2020, 19, 2131–2167.
3. Almars, A.M.; Almaliki, M.; Noor, T.H.; Alwateer, M.M.; Atlam, E. HANN: Hybrid Attention Neural Network for Detecting Covid-19 Related Rumors. IEEE Access 2022, 10, 12334–12344.
4. Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web, Montreal, QC, Canada, 11–15 April 2016.
5. Waseem, Z.; Hovy, D. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In Proceedings of the NAACL-HLT, San Diego, CA, USA, 12–17 June 2016; pp. 88–93.
6. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. HAutomated Hate Speech Detection and the Problem of Offensive Language. In Proceedings of the International AAAI Conference on Web and Social Media, Montreal, QC, Canada, 15–18 July 2017; pp. 88–93.
7. Fortuna, P.; Nunes, S. A Survey on Automatic Detection of Hate Speech in Text. ACM Comput. Surv. 2018, 51, 1–30.
8. Sharma, S.; Agrawal, S.; Shrivastava, M. Degree based classification of harmful speech using twitter data. arXiv 2018, arXiv:1806.04197.
9. Almars, A.M. Attention-based Bi-LSTM model for Arabic depression classification. Comput. Mater. Contin. 2022, 71, 3091–3106.
10. Djuric, N.; Zhou, J.; Morris, R.; Grbovic, M.; Radosavljevic, V.; Bhamidipati, N. Hate Speech Detection with Comment Embeddings. In Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 1–6.
11. Badjatiya, P.; Gupta, S.; Gupta, M.; Varma, V. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, 3–7 April 2017; pp. 1–6.
12. Gitari, N.D.; Zhang, Z.; Damien, H.; Long, J. A Lexicon-based Approach for Hate Speech Detection. Int. J. Multimed. Ubiquitous Eng. 2015, 10, 215–230.
13. Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; Weber, I. Analyzing the Targets of Hate in Online Social Media. Proc. Int. AAAI Conf. Web Soc. Media 2021, 10, 687–690.
14. Kwok, I.; Wang, Y. Locate the Hate: Detecting Tweets against Blacks. Proc. AAAI Conf. Artif. Intell. 2013, 27, 1621–1622.
15. Mubarak, H.; Darwish, K.; Magdy, W. Abusive Language Detection on Arabic Social Media. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, 4 August 2017; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 1–6.
16. Mohammad, A.S. Mother tongue versus Arabic: The post-independence Eritrean language policy debate. J. Multiling. Multicult. Dev. 2016, 37, 523–535.
17. Darwish, K.; Magdy, W.; Mourad, A. Language Processing for Arabic Microblog Retrieval. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, HI, USA, 29 October–2 November 2012; Association for Computing Machinery: New York, NY, USA, 2012; pp. 2427–2430.
18. Burnap, P.; Williams, M.L. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. Policy Internet 2015, 7, 223–242.
19. Magdy, W.; Darwish, K.; Weber, I. #FailedRevolutions: Using Twitter to study the antecedents of ISIS support. First Monday 2016.
20. Kaati, L.; Omer, E.; Prucha, N.; Shrestha, A. Detecting Multipliers of Jihadism on Twitter. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; pp. 1–6.
21. Atlam, E.S.; Fuketa, M.; Morita, K.; Aoe, J.-i. Similarity measurement using term negative weight and its application to word similarity. Inf. Process. Manag. 2000, 36, 717–736.
22. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv 2019, arXiv:1910.01108.
23. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2018, arXiv:1810.04805.

24. Bonifazi, G.; Corradini, E.; Ursino, D.; Virgili, L. New Approaches to Extract Information From Posts on COVID-19 Published on Reddit. Int. J. Inf. Technol. Decis. Mak. 2022, 21, 1385–1431.
25. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. arXiv 2019, arXiv:1907.11692.
26. Chouikhi, H.; Chniter, H.; Jarray, F. Stacking BERT based Models for Arabic Sentiment Analysis. In Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Valletta, Malta, 25–27 October 2021; pp. 1–6.
27. Zhang, C.; Abdul-Mageed, M. No Army, No Navy: BERT Semi-Supervised Learning of Arabic Dialects. In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 28 July–2 August 2019; pp. 1–6.
28. Alammary, A.S. BERT Models for Arabic Text Classification: A Systematic Review. Appl. Sci. 2022, 12, 5720.